USAID
FROM THE AMERICAN PEOPLE


# PERFORMANCE & IMPACT EVALUATION (P&IE)

## Semi-Annual Report, April 2015

**30 APRIL 2015**

This publication was produced at the request of the United States Agency for International Development. It was prepared independently by NORC at the University of Chicago.

# PERFORMANCE & IMPACT EVALUATION (P&IE) SEMI-ANNUAL REPORT, APRIL 2015

30 April 2015

PN 7384; USAID Contract N0: AID-617-C-12-00006

**PRESENTED TO:**

USAID/Uganda

Joseph Mwangi

**PRESENTED BY:**

NORC at the University of Chicago

Jeffrey Telgarsky

Executive Vice President, Research

4350 East-West Highway, 8th Floor

Bethesda, MD 20814

Telephone: (301) 634-9413

Fax: (301) 634-9301

**DISCLAIMER**

# CONTENTS

NORC at the University of Chicago (NORC), in partnership with the Panagora Group, is pleased to submit to USAID/Uganda this Semi-Annual Report recording progress made on the P&IE project between November 1, 2014 and April 30, 2015.

# HIGHLIGHTS

During this reporting period, the NORC/Panagora team:

- Completed a data quality review of the EGRA Cluster 1 Round 3 and Cluster 2 Round 2 datasets. This data, which will be used, along with Cluster 1 and Cluster 2 baseline and Cluster 1 Round 2 data, to measure the impact of SHRP's Reading (Result 1) interventions was collected by the IP in October-November 2014. NORC received the dataset in January 2015, and reviewed it for quality and consistency with the previous rounds of Cluster 1 and 2 datasets in preparation for the impact analysis.
- Completed the second annual impact analysis for Result 1 using the baseline and Round 3 data from Cluster 1 schools and baseline and Round 2 data from Cluster 2 schools. Prepared and submitted to USAID the two Impact Evaluation Reports (one for Cluster 1 and a separate report for Cluster 2), documenting progress in outcome indicators after two years of SHRP interventions in Cluster 1 schools and after one year of the interventions in Cluster 2 schools[1].
- Conducted numerous data quality assessment (DQA) tasks related to the Cluster 3 EGRA for which baseline data collection took place in February 2015:
    - Reviewed EGRA instruments and provided feedback to the IP, both from the perspective of NORC's role as evaluator and data quality reviewer. This included reviewing and commenting on changes to the assessment instrument and sample, taking care to ensure that these changes would not have significant negative implications for the impact evaluation.
    - Reviewed enumerator training manuals for the Cluster 3 baseline data collection.
    - Participated in enumerator training for the Cluster 3 baseline data collection, which took place in February 2015.
    - Travelled to four primary schools in three Cluster 3 districts (Pallisa, Iganga, and Kibuki) to observe data collection activities for the EGRA data collections.
    - Addressed changes to the Result 2 KAP 2 school sample, which will have implications for the scope of the Result 2 impact evaluation.
- Continued the ongoing process of third party monitoring and performance feedback with meetings and activities, based on a monthly calendar and an events and assignments tracker to ensure comprehensive monitoring.
    - Continued to observe RTI/SHRP implementation activities, including leadership and management trainings, school support supervision exercises, and EGRA Assessors trainings. During this period 16 events were observed.

[1] Since only baseline data is available for the School Health (Result 2) at the time of this report, the impact analysis for Result 2 will be presented in the third annual impact evaluation report after follow-up data has been collected in October 2015.

  - Continued to record, using our observation tools, information and observations on each activity for the upcoming final performance evaluation; and noted appreciative and constructive comments to provide as monthly performance feedback to RTI within the context of the CLA process.
  - Continued to conduct performance feedback meetings, on a monthly or bi-monthly basis, with SHRP Chief of Party and M&E Director, as well as key Results 1 and 2 team members based on observations from SHRP events and activities. During this reporting period, we held two performance feedback sessions (November, February), with RTI/SHRP.
  - Initiated a new monthly meeting between in-country staff and SHRP R1 Reading and R2 Health leads to allow for a more holistic understanding and broader view of overall SHRP progress and achievements.
- Conducted a thorough review and revision of the Mid-Term SHRP Performance Evaluation Report, addressing USAID and RTI comments, and submitted a revised version to USAID on April 3, 2015.
- Delivered a presentation titled "Impact evaluation of a mother-tongue based early grade reading program in Uganda" at the Comparative and International Education Society (CIES) conference in Washington, DC on March 11. The presentation covered the methodology and preliminary results from the USAID Uganda Performance and Impact Evaluation of the School Health and Reading Program.

# DESCRIPTION OF ACTIVITIES

## A. IMPACT EVALUATION ACTIVITIES

### Sample Issues around EGRA Cluster 3 Baseline Data Collection

The P&IE Evaluation Expert continued to weigh in on sampling issues as they arose for Cluster 3. In particular, during this reporting period, two sample concerns emerged and were resolved.

(1) The Cluster 3 sample that NORC received from RTI in February included only 1 control CCT per language. The reason for this was that all other non-treated clusters in the Cluster 3 districts were slated to receive treatment in 2016 so that SHRP could meet its student targets. SHRP staff explained to NORC that, for various reasons (security situation, ongoing MoES programs, budget constraints, etc.), expanding to new districts with Cluster 3 languages was not an option.

NORC expressed concern about this sampling change. Namely, that having only 1 control CCT in each district poses a risk to the impact evaluation because if for any reason that control CCT is different (for instance if it is a wealthier area, or if local authorities are much worse or better than treatment CCTs, etc.), the treatment and control will not be sufficiently similar to support an RCT. As such, we requested that, at a minimum, RTI add at least one more control CCT per district (even though more than one CCT would have been preferable). We also requested that all schools in the added control CCT remain untouched for the period of the study.

Ultimately, after lengthy email discussions, it was agreed that RTI would add 1 more CCT in each Cluster 3 district. However, due to the limited number of schools in these districts and the need to reach a specified target of students, SHRP would intervene in all but the control schools in these CCTs in 2016.

All parties agreed to this solution, while acknowledging the risks of contamination in control CCTs where some schools receive the SHRP interventions. Contamination of control schools could potentially lead to underestimation of the impact of SHRP interventions.

(2) RTI opted to drop English-only schools from the C3 sample because they realized that the program materials are not consistent with the needs of such schools. However, this exclusion applies only to the C3 sample; no schools will be excluded from the C1 and C2 samples. NORC accepted this change.

## Sample Issues around Midline KAP Data Collection (scheduled for November 2015)

In February, NORC learned from WorldEd that, at NARC's request, they were dropping the KAP2 schools from the mid-term data collection and beyond because adequate approvals from NARC had not been garnered for the new districts and schools in KAP2. As a result, future data collections (midline and endline) would only include KAP1 schools.

This decision has implications for the impact evaluation. The KAP2 data collection included schools from 4 new districts, which will no longer be part of the impact evaluation. Moreover, it also included boarding schools from the original KAP1 sample that could not be included because of missing consent forms during KAP1 data collection. Without the boarding schools in the KAP2 school sample, we will not be able to assess the impact of Result 2 activities on boarding schools.

In the absence of the KAP2 sample, the impact evaluation will only be able to assess the impact of the Result 2 interventions in secondary day schools and not secondary boarding schools. Indeed, although KAP1 data collection included boarding schools, the boarding school sample visited during KAP1 was not representative of that population. The boarding school students interviewed were not a random sample of students since only those with consent forms or those who were over 18 and could give consent themselves were included. The sample of schools was also not representative. As shown in the table below, partly boarding schools make up 39% of the population of schools and full boarding schools make up 15% of the population of schools; however in our KAP1 data, they make up 20% and 3.7% of the sample of secondary schools, respectively. Given that neither the secondary school nor the student samples are representative, we will not analyze post-primary boarding schools, and will restrict our analysis of SHRP impacts on secondary schools to the day schools only.

| | Population | KAP 1 DATA |
|---|---|---|
| Day Schools | 46% | 75.80% |
| Partly boarding | 39% | 20.40% |
| Full boarding | 15% | 3.70% |

## Addressing Changes to EGRA Cluster 3 Data Collection Tools

In January, NORC learned that SHRP was considering revising the content of the EGRA tools by removing some subtasks and adding new ones, as is explained in the following table.

| Subtask | Action | Justification |
|---|---|---|
| English Letter Sound Knowledge | Keep | Children are not taught to read in P1 English. Yet, this measure can capture some individual differences at baseline (from PreK or other environmental influences). |

| Subtask | Action | Justification |
|---|---|---|
| English Segmenting | Remove | Removing the segmenting subtask will increase validity of the segmenting results in the other language. English and local language segmenting are similar but the sound unit differs (syllable versus sound). This could confuse. |
| English Nonword Decoding | remove | Remove to reduce the length of the assessment. Plus, this subtask assesses decoding. P1 learners have not been taught to decode nor will they be taught to decode in P1. |
| English Oral Passage Reading With Comprehension | Remove from P1. Include in P2. | Children do not learn to read connected text in English in P1. P1 letter sounds can be used to inform P2 English reading skills. |
| English Vocabulary | Keep | The focus of P1 English is to develop oral language skills. This assessment measures receptive language reliably. |
| Local Language Letter Sound Knowledge | Keep | Letter sound knowledge supports word recognition. |
| Local Language Segmenting | Keep (remove English) | Remove one of the segmenting subtasks will increase validity of the segmenting results in the other language. English and local language segmenting are similar but the sound unit differs (syllable versus sound). This could confuse. |
| Local Language Nonword Decoding | Remove | Remove to reduce the length of the assessment. And this task is primarily diagnostic to compare decoding skills to word recognition ability (measured via passage reading). Children in this sample do not have minimal word recognition skills so this task does not offer information. |
| Local Language Oral Passage Reading with Comprehension | Keep | Authentic reading task. The P1 curriculum teaches children to read connected text in the local language. |
| Local Language Listening Comprehension | Keep | This is a diagnostic task. It helps to explain results on passage comprehension |
| Orientation of Print | Add | For children in our sample, it will often be the first time they learn directionality, finger sweeping and other print concepts. |
| Letter Writing | Add | This subtask measures alphabetic principal. With five items and scored qualitatively, it is child-friendly. |

Source : SHRP

The main justification for modifying the EGRA tools was to shorten the assessment and only include subtasks that were more directly relevant to the skills acquired by children in P1. SHRP also confirmed that the non-word decoding tasks in English and local language, as well as the segmenting task in English would be re-introduced in the P2 EGRA tool.

NORC agreed to the proposed changes by SHRP. We also note that while non-word decoding is a difficult task for students, it is a good predictor of ability to decode real words but agreed to remove the task to reduce the length of the assessment. We welcomed the addition of the Orientation of Print and Letter Writing subtasks as these were also recommended by NORC during the Evaluation Inception phase.

## Data Quality Assessment

During this reporting period, P&IE staff engaged in various data quality assessment (DQA) tasks related to the Cluster 1 Round 3 and Cluster 2 Round 2 EGRA datasets and the Cluster 3 EGRA data collection activities. They included the following:

- Conducted a data quality review of the Cluster 1 Round 3 and Cluster 2 Round 2 EGRA datasets. NORC staff noted a few issues with the data – missing values, problematic or missing value and/or variable labels, etc., but overall, we found the data to be of high quality. NORC staff communicated these quality issues to RTI and received feedback and an adjusted dataset, which we used for the impact analysis.
- Reviewed and provided feedback on all data collection instruments (EGRA and learner context instruments, teacher/head teacher survey, classroom observation tool, school inventory) for the Result 1, Cluster 3 EGRA data collection and provided written feedback to the SHRP team in January 2015 prior to the enumerators training. We noted that the tools were largely similar to previous rounds (except for the changes noted above) and the main problems noted were with skip patterns, formatting and editorial issues, rather than problems with substantive content.
- Participated in the enumerator training for the Cluster 3 baseline data collection. The P&IE team's Literacy Expert, Mark Lynd, travelled to Uganda on February 9-21 to participate in the enumerator and supervisor trainings and field testing of the instrument. NORC staff observed that the IP had implemented some of the recommendations we had made in previous rounds, such as giving feedback to assessors immediately after the IRRs. However, we also noted that some concerns we had previously alerted the IP to were still present, such as the use of clipped sounds and only accepting one sound per letter, even though some letters clearly have several associated sounds. We also noted some new concerns. For issues/problems with quick fixes, which in our opinion would contribute to the quality of the instruments, we provided immediate verbal feedback to the IP. Other (bigger) issues that require more complex solutions have been documented in Annex 1 of this report and will be shared with the IP, so they can be taken into consideration for the next round of EGRA data collection.
- Conducted field observations in Pallisa, Iganga, and Kibuki districts on February 24-27, 2015. NORC's Resident Evaluation Manager travelled to the field to observe field work for the EGRA data collection in four primary schools in the 3 districts. Observations resulting from these field visits are presented in Annex 2; these observations were shared with the IP in NORC/Panagora's March 2015 Feedback Memo to RTI.

## Impact Analysis and Impact Evaluation Report

P&IE staff conducted extensive analysis of EGRA data from Cluster 1 and Cluster 2 schools for the second annual impact evaluation of Result 1 activities conducted under SHRP. This included data cleaning and preparation of datasets for analysis, and subsequent analysis consisting of ordinary least squares multinomial regressions to evaluate the impact of SHRP on early reading skills. Different models were used to check the robustness of the results. The approach follows very closely the original analysis plans outlined in the approved evaluation design. We prepared and submitted to USAID two Impact Evaluation Reports, for Cluster 1 and Cluster 2 respectively, documenting the impact of SHRP in outcome indicators in Cluster 1 after two years of implementation and in Cluster 2 schools after one year of implementation.

# B. PERFORMANCE EVALUATION AND CLA ACTIVITIES

## Systematic observation and documentation of project implementation activities as inputs for the performance evaluations.

The P&IE team, led by subcontractor Panagora Group, continued to implement our processes for systematically monitoring and documenting SHRP implementation activities, and our system for carrying out the CLA elements of our contract and providing RTI with performance feedback.

During this period, we continued to follow a sequence of information collection, reporting, review, and feedback that includes the following:

- A monthly meeting between in-country P&IE staff (Resident Evaluation Manager and Sr. HIV/AIDS Specialist) to review the prior month's work and determine content of the monthly report and the performance feedback memo to RTI based on information from the observation reports
- Preparation and submission of a monthly activities report and draft SHRP performance feedback memo by P&IE country staff to U.S.-based team
- Full P&IE team meeting to discuss performance feedback memo and monthly in-country activities (U.S.-based and in-country staff)
- Bi-monthly performance feedback meeting with SHRP project staff (at the request of SHRP COP, this meeting was shifted from a monthly to a bi-monthly meeting)
- Performance Evaluation/CLA coordination meeting (Panagora, in-country staff)

The memos, meetings and feedback activities listed above have the end goal of accurately documenting the P&IE team's observations of implementation activities and providing appreciative and constructive feedback to the SHRP team, so they can use our observations and suggestions to improve implementation in real time. The memos will also serve as input into the Final Performance Evaluation.

During this reporting period, our in-country staff attended and observed the following meetings, events, and trainings, and prepared a report on each of them using the appropriate observation and monitoring tool. Each of the reports provided appreciative and constructive observations that were collated and shared with RTI in monthly performance feedback memos.

- EGRA Main Assessors Training, observed September 29 – October 3, 2014
- EGRA Field Data collection, observed October 6 – 8, 2014 in Lumasaba, October 14 – 15, 2014 in Luganda, October 22 – 23, 2014 in Runyoro-Rutooro
- Early Grade Reading Master Training on P3 materials, December 9, 10, and 12, 2014, at Shimon PTC
- Training of Trainers on P2/P3 Early Grade Reading materials: P2/P3 at Ngora PTC on December 16, 2014
- Training of Trainers on P2 Early Grade Reading materials at Nyondo PTC on December 17, 2014
- Training of Trainers on P1 Early Grade Reading materials at Bishop Willis PTC on December 18, 2014
- Teachers Early Grade Reading training on C1 P1/P2/P3 materials at Shimon PTC on January 6 – 7, 2015

- Teachers Early Grade Reading training on C1 P1/P2/P3 materials at Kabulasoke PTC on January 13, 2015
- Teachers Early Grade Reading training on C3 P1 materials at Bwera PTC on January 21, 2015
- Teachers Early Grade Reading training on C2 P1/P2 materials at Canon Apollo Kabarole PTC on January 22, 2015
- C3 EGRA Assessors Training, observed February 10 – 19, 2015
- C3 EGRA Supervisors Training, observed February 20, 2015
- C3 Field Data Collection, observed February 24 – 27, 2015
- Radio Talk Shows Preparatory Meeting to promote community/public knowledge on the value of early grade reading, observed February 11, 2015
- Materials Development Workshop for C1 P4, observed March 16 – 17, 2015
- Monitoring and Support Supervision Activities for R1 and R2: Bushenyi district March 10 – 12, Wakiso district March 10 – 11, Lira district March 17 – 18, Arua district March 24 – 26

## Collaboration, Learning and Adaptation (CLA) Activities

To implement the CLA component of the P&IE contract, we continued to provide performance feedback to RTI on a monthly basis, with both appreciative and constructive feedback, focusing on elements of performance where real-time feedback will help to strengthen performance and lead to optimal outcomes.

The performance feedback continues to be drawn directly from the reports of meetings, events, and activities observed by P&IE in-country staff. Each observation tool includes a section to note and record both appreciative and constructive feedback. Our Resident Evaluation Specialist and Senior HIV/AIDS Evaluator continued to collate these comments into a monthly performance feedback memo. Panagora continued to finalize the memo, integrating comments from the full P&IE team, provide the memo in advance to RTI, and lead the feedback session with RTI's leadership and the full P&IE team participating.

Five months of performance were covered during the reporting period; October, December, January, February, and March. In November, the Resident Evaluation Manager was participating in a training and the Senior HIV/AIDS Evaluation Specialist was out on sick leave. RTI continues to express its appreciation for the value of the feedback memos and the subsequent discussions, which have given them access to insights that allow them to improve their performance in real time. We are pleased that this exercise continues to be regarded by RTI as a valued and welcome opportunity to improve performance and results achievement as a part of regular implementation.

Annex 3 presents an illustrative example of appreciative and constructive performance feedback provided during the reporting period, including RTI's response (in italics).

## Midterm Performance Evaluation Planning

During this reporting period, our subcontractor leading performance evaluation work under P&IE, Panagora Group, conducted a thorough review and revision of the Mid-Term Performance Evaluation Report, which was originally submitted to USAID on September 30, 2014, and was re-submitted on April 3, 2015. Panagora Group integrated comments from USAID and RTI.

## C. RISKS TO THE IMPACT EVALUATION

Below we present an assessment of risks/challenges to the impact evaluation design that emerged during this reporting period. Challenges identified prior to the current reporting period, which were presented in previous semi-annual reports, are listed in Annex 4.

1. Result 1: Possible contamination of controls in Cluster 3. Because SHRP team is not planning to expand SHRP implementation to additional districts for Cluster 3, they will implement Result 1 activities in control CCTs in Cluster 3 districts starting in 2016 in order to meet ambitious targets. However, they plan to exclude the control schools within the control CCTs which were selected for the EGRA data collection (and evaluation sample) and intervene only in the schools from control CCTs that are not part of the evaluation sample.

   We faced an identical situation with Cluster 1 CCTs. Strict exclusion of control schools from treatment is critical for the integrity of the impact evaluation design. While SHRP staff has assured us that control schools will be excluded from the Result 1 interventions, we are nonetheless concerned by the possibility of contamination through CCTs or spillover of materials. Any contamination of the control schools will lead to underestimation of the effects of the SHRP Result 1 interventions.

2. Result 2: Exclusion of secondary boarding schools from Result 2 impact evaluation: In February, NORC learned from WorldEd that, at NARC's request, they were dropping the KAP2 schools from the mid-term data collection and beyond because adequate approvals from NARC had not been garnered for the new districts and schools in KAP2. As a result, future data collections (midline and endline) would only include KAP1 schools.

   This decision has implications for the impact evaluation. The KAP2 data collection included schools from 4 new districts, which will no longer be part of the impact evaluation. More importantly, it also included boarding schools from the original KAP1 sample that could not be included due to consent issues during KAP1 data collection. Without the boarding schools in the KAP2 school sample, we will not be able to assess the impact of Result 2 activities on boarding schools.

   In the absence of the KAP2 sample, the impact evaluation will only be able to assess the impact of the Result 2 interventions in secondary day schools. The boarding school sample visited during KAP1 is not representative of that population. The boarding school students interviewed were not a random sample of students since only those with consent forms or those who were over 18 and could give consent themselves were included. The sample of schools was also not representative. Partly boarding schools make up 39% of the population of schools and full boarding schools make up 15% of the population of schools; however in our KAP1 data, they make up 20% and 3.7% of the sample of secondary schools, respectively. Given that neither the secondary school nor the student samples are representative, we will not analyze post-primary boarding schools, and will restrict our analysis of SHRP impacts on secondary schools to the day schools only.

# ANNEX 1: FEEDBACK FROM OBSERVATION OF ENUMERATOR TRAINING AND PILOT TEST FOR CLUSTER 3 EGRA DATA COLLECTION

## INTRODUCTION

The School Health and Reading Program (SHRP) funded by USAID/Uganda conducted a workshop in which assessors were trained to conduct a baseline Early Grade Reading Assessment (EGRA) in Cluster 3 schools. This report presents a summary of that workshop, including effective practices observed during the training as well as questions, concerns, and recommendations.

## OVERVIEW OF THE WORKSHOP

The workshop was conducted from Tuesday February 10 to Friday February 20. Dr. Mark Lynd from School-to-School International and Ms. Evelyn Namubiru, Resident Evaluation Manager for NORC, attended the workshop in the capacity of Quality Assurance monitors with the purpose of observing and providing feedback on the quality of the training and baseline tools. Dr. Lynd observed the workshop from February 10-19, and Ms. Namubiru from February 10-20. The workshop was facilitated by the following:

- Tracy Brunette, M&E Director, SHRP
- Rehemah Nabachwa, M&E Specialist, SHRP
- Peter Muyingo, M&E officer, SHRP

And the following DQAs:

- Lydia Nakijjoba, DQA
- Stella Kambugu, DQA
- Deborah Nakyejwe, DQA
- Rosette Kyalisiima, DQA

Also attending at various points were the following:

- Saeeda Prew, Chief of Party, SHRP
- Derek Nkata Deputy Chief of Party, SHRP
- Daniel Nkaada, Commissioner MOE Basic Education (he opened the training)
- Four officers from UNEB

The room was arranged with 57 participants grouped by language at 4 large tables. The languages and locations represented were as follows:

| Language | Number of assessors | Region | District |
|---|---|---|---|
| Lugwere | 14 | Mid -Eastern | Budaka, Pallisa, Kibuku |
| Nkarimojong | 14 | North East | Nakapiripirit, Napak, Moroto |
| Lukhonzo | 14 | Mid-Eastern | Kasese |
| Lusoga | 15 | East Central | Iganga and Kamuli |
| **Total** | **57** | | |

The following is a summary of that workshop, including effective practices observed during the training as well as questions, concerns, and recommendations. (Note: The Supervisor training was attended by Evelyn Namubiru; comments on that training are provided by Ms. Namubiru.)

## GOOD PRACTICES

### Over-recruitment of assessors.

For each language group, 14 assessors participated in training, the best 12 of whom were retained for actual data collection.

### Effective workshop design.

On balance, workshop content and process were very good.

a. Adequate time allotted – 9 days total, including 1 day school practice.
b. Effective management and use of groups as this was a large group to train.
c. Knowledgeable, dynamic facilitators. Most of the training was conducted by DQAs, who were extremely knowledgeable in the content of the EGRA instrument and its use, and dynamic in their delivery and facilitation of group sessions.
d. Hands-on, active learning. Overall, assessors were highly engaged in practice activities and had substantial time with facilitators, both in pair practice and in guided practice (group) sessions.
e. Assessors spent substantial time practicing EGRA, using the tablets, and discussing subtasks, items, marking procedures, and questionnaire questions.
f. Effective sessions included the following features:
    i. Visual support. Facilitators made effective use of the video projector for PowerPoints and projection of Tangerine screens to highlight how to mark responses. Facilitators also made effective use of video through the video-recorded presentation of simulations of EGRA subtasks and a video of Lydia, DQA, trainer, and literacy expert, pronouncing each letter of the alphabet while displaying a the corresponding letter in writing. The IRR visual support was particularly effective (see next point).
    ii. Inter-rater reliability. The use of color-coded "gold standard" results to highlight reliability (consistency of how items were marked) effectively facilitated discussions about items that were the most problematic and discuss strategies for improving (e.g., skipped items should be marked as incorrect), as well as facilitators' own performance. (Note that NORC had previously suggested that in these simulations, the "child" being assessed should use a script in order to be able to recall the type of errors being made – e.g., skipped, letter name instead of sound – for the IRR discussion afterward. SHRP adopted this recommendation in this workshop.)
    iii. Feedback to assessors. Much of the workshop consisted of "pair practice" during which assessors practiced various subtasks with a colleague while DQAs and project staff observed, then gave feedback. Also used were "guided practice" in which DQAs worked with assessors in their language groups to discuss problematic areas. Assessors also received feedback on Wednesday when they took a pop quiz and were given their results. Additionally, the IRR activity provided assessors with information on their individual performance implementing EGRA, enabling them to focus on the specific items within those subtasks on which they had scored below "the gold standard" of 90% or more.
    iv. Use of video projections in which Tangerine windows were projected so plenary corrections could be made during a simulation. This was effective for showing when something should be marked right or wrong.

### Effective field practice design

a. Well organized: each assessor had the opportunity to practice conducting EGRA five to six times - three P1 and three P3 learners - in real conditions.
b. Debrief afterward was thorough, constructive, and edifying, targeting key problems we observed during the practice.
c. Session following school practice was effective back in Kampala. Results from practice were shared with assessors to highlight what they had done and how they could improve (e.g., too long on certain tasks). In language groups, the DQA also discussed observations specific to individual assessors and gave them support and feedback to improve.

### Supervisor training

Supervisors were involved in preparation of field materials that will be required to use when they reach schools. They were guided on writing details of schools to be visited and data collection activities that would take place at each school. They also prepared learner sampling cards.

## QUESTIONS

During the workshop, NORC met with RTI and shared some questions. Below is the list of questions with RTI responses.

a. Incorrect vs. no response. Some items require the assessor to choose between incorrect and no response; other items combine these two. The distinction was highlighted as important in the training, and was included as one of the items on the quiz on Wednesday. Yet to date, the responses have not been analyzed separately, so why is this distinction being highlighted? RTI indicated that they will review whether this distinction should be emphasized in future trainings. However, we also understand that for timed subtasks, an analysis of items attempted is not possible if the two are not differentiated.
b. Different oral comprehension stories. Each of the four language EGRAs has a different oral comprehension story. No provisions have been made to equate the difficulty of these stories. RTI reported that these stories are different because local language experts opted to develop the stories independently. RTI support staff at HQ will examine this situation (see 2a below).
c. Assessor performance. Is there a way to rate assessors' performance with students? RTI reported that they use multiple tools (at one point, eight was stated), including the IRR observations, observation checklists, and the quiz given on Day 2. It is unclear if these assessments are recorded or used in a systematic way. When assessors were deselected, the next day, RTI announced that the decisions had been made on the basis of IRR scores as well as other factors such as behavior, demeanor with students, etc.
d. Equating. Are EGRAs different from one administration to the next? RTI response indicated that with cohorts 1 and 2, equating was conducted.[2] With cohort 3, the tasks and task order will remain the same from one administration to the next. For the letter sound subtask, the same letters will be used but rearranged. For cohorts 1 and 2, the timed reading stories were

- [2] The Data Quality Assessment Report (NORC, October 2013) notes that "(t)he EGRA tools remained the same as the ones used during Round 1, except for two major changes: (1) items were re-randomized in the Letter Sound Knowledge and Nonword Decoding sections, and (2) the reading passage for the Oral Reading Passage in local language was changed, using a passage that had been equated with the Round 1 passage during the testing phase of EGRA (the listening comprehension story changed only slightly as some proper names were modified, and the English story did not change). NORC agrees with these changes." The comparability of the cohort 3 tools should be confirmed as well.

different between languages, and were equated to ensure comparable difficulty; however, for cohort 3, this was not done due to time constraints. Since timed reading texts were not equated, RTI will statistically adjust scores across languages from the baseline to ensure comparability although details of how this will be done have not been shared. For the listening comprehension passages, RTI will use the same stories across languages but change key words – e.g., mangoes to bananas, character names. We believe that the listening comprehension adjustments are acceptable. However, if possible, texts between languages should be of similar difficulty and some equation process should ideally occur at the time of development. Although we understand that oral reading fluency in terms of correct words per minute should not be compared across languages, we think it would still be worthwhile to ensure that stories across languages are of equivalent difficulty.

## CONCERNS & RECOMMENDATIONS

### Workshop

a. Organization/logistics
   i. At times, the projected text was too small to read from the back of the room. Recommend: Ensure that all projected material is visible from the back of the training room.
   ii. Sometimes trainers could not be heard or spoke so forcefully into the mike that s, th, sh all sounded like white noise, potentially compromising sound recognition exercises. Recommend: Train facilitators to use the microphone properly.
   iii. One literacy expert (Lydia) was in attendance throughout the workshop, and according to RTI, two or three language experts were available on Day 1 and 2 of the workshop (all except Ngakarimajong). Greater availability of these experts would have helped to inform discussions concerning correct pronunciation, high-frequency letters, reading issues, etc., facilitators deferred to the DQAs and project personnel. Recommend: Increase the availability of experts.
b. Content
   i. PowerPoint: Overview of EGRA slide: What does EGRA measure? Fluency: Slide says EGRA measures reading speed **and proper expression**. Recommend: Delete "and proper expression" since EGRA doesn't measure this.
   ii. PowerPoint: Overview of EGRA slide: What does EGRA measure? Comprehension. Slide says EGRA measures whether students understand **and actively engage in text**. Recommend: Since EGRA doesn't measure a child's active engagement in text, delete this.

c. Instructional approach
   i. Difficulty of sounds. Consistent marking of sound production was one of the biggest challenges noted in this workshop. Pronunciations can vary by language group and region, and sometimes conditions made it difficult to clearly discern sounds (see "use of microphones" above. Recommend: Provide assessors with references for the pronunciation of sounds for continued reinforcement of standard, "correct" pronunciations of letters – e.g., load the pronunciation video on their tablets or send to their phones. Note that this recommendation was taken up by RTI during the training: assessors were asked to download the pronunciation video onto their tablets.
   ii. The video presentations of subtasks typically consisted of simulations with facilitators, never with real children in real contexts. While there is value in showing simulations with facilitators and peers, additional value could be gained by showing videos of real-life situations. Recommend: Use real-life contexts in videos.
   iii. When reviewing IRR results or simulations of subtasks, it would be helpful if the assessors could review what was asked and how the "child" responded so they can think about why they had made a mistake. Video stop-action is well suited for this. Recommend that the IRR sessions be videotaped, then shown when presenting IRR results to help assessors understand what actually happened. Recommend: Use stop-action in videos.
   iv. The effectiveness of IRR could be improved by reducing the time between IRR simulations and presentation of results – e.g., if possible, by doing one subtask, then uploading, analyzing during break, and discussing the results with assessors the same day. Recommend: Present IRR results as soon after a simulation as possible. RTI response: Agreed that this is a good idea but the process of uploading data, cleaning, analyzing, and preparing slides takes several hours, often resulting in people working late into the night. Our note: If the statistician could be present at the workshop, this turnaround time might be reduced.
   v. When giving examples of acceptable responses, some counterexamples could help with understanding. For example, when showing which letters for the writing subtask are acceptable, also display which ones are not acceptable. Recommend: Provide counterexamples where relevant.
   vi. The workshop was characterized by a presentation/ practice/ discussion format that effectively involved assessors in discussions. The workshop also made extensive use of hands-on activities in which assessors simulated the use of EGRA with paper forms and tablets, followed by pair practice in which assessors practiced the skill under the observation of a DQA or staff member, then a discussion concerning how the process went. While this model can be effective, presentations mostly consisted of PowerPoint presentations with long descriptions of how to administer EGRA. Sometimes, support was provided via video or simulations, but most of the time, facilitators talked at length about procedures, flipping from one slide to the next, one facilitator asking repeatedly "Are you with me?" – a strategy that doesn't ensure attention or interest.
   vii. The use of more experiential methods would improve training quality. One example is the experiential learning cycle, in which assessors first *experience* new content (by doing or observing it), then *reflect* on it (describing, asking questions, hypothesizing), *apply it* (practice it, discuss its meaning, how it works, why it's important, how they should do it), and finally, *plan* for use of the new content in their work (next steps, consideration of implications or possible difficulties). In other words, first immerse assessors in an experience that they can reflect on, then reinforce with content, rather than the reverse. For example, rather than presenting a PowerPoint and describing the letter sound knowledge subtask question by question, start with a demonstration (facilitators or assessors) then discuss what assessors saw, what was done right, what was done wrong, and what lessons were learned. Then

present guidelines: at this point, the assessors will be more able to understand and absorb them. Recommend: Make greater use of experiential methods.

viii. Though agendas were provided at the beginning of each day, no learning objectives were presented for any session (e.g., by the end of this session, assessors will be able to…), nor were assessors given opportunities for demonstrate of new knowledge or skills to show that the session objectives had been met. Recommend: Begin sessions by explicitly introducing the learning objectives for each exercise; by ensuring that assessors understand the objectives and targeted learning outcomes (content or skill) of each practice session; and by giving assessors the opportunity to evaluate whether the learning objectives had been met. Recommend: Make instructional goals explicit.

ix. No evaluations of the training were conducted from Day 1 through 9 (Feb. 19). Recommend: On a periodic basis (e.g., at the end of sessions or the end of the day), evaluations are conducted so that assessors rate what they learned most, what is still difficult, and the quality of the training. Facilitators should also evaluate learning in a systematic and regular way; the quiz was a good example.

## Structure of EGRA

Test-retest problem. The Orientation to print story is the same as the Timed reading passage. Though it probably will not make a significant difference, these passages should ideally be different. Recommend: provide different passages for each task.

## EGRA tool

a. Building rapport section - Translation issues. Some assessors reported that translations can be inappropriate. For example, when asked "Would you like to answer these questions?" one the child declined, whereas when the assessor said "Would you like to play/participate?" the child agreed. Recommend: Check translations to ensure all invite the child to participate, not just answer questions.

b. Letter-sound task:
   i. Incomplete letter coverage. Some of the four Ugandan alphabets include combined letters outside of the Roman alphabet. For example, Lukhonzo contains gh, gy, ky, lh, nd, ng, ngy, nt, ny, nz, th, teachers. However, none of these appear in the letter sound recognition subtask. Recommend: Ensure that all letters are included in the letter sound subtask at least once. RTI response: Some languages were changing while EGRA was being developed (e.g., Ateso) so the decision was taken to keep the core letters only in each language.
   ii. Making clipped sounds only as correct. For example, when asked to pronounce the sound of the letter "b," responses are marked correct if students give the clipped version - /b/ - and incorrect if they aspirate – e.g., /buh/. This assessment practice biases the results in favor of treatment schools, where teachers are trained in the "correct" way to assess this skill, even if /buh/ also shows that the child understands the letter sound (as opposed to the letter name for example). Note that this issue was discussed in February 2014 (see the October 2014 DQA report). We are repeating this observation because we believe this practice poses a threat to the validity of the test that should be mentioned in the report.
   iii. The use of only one sound per letter. On the letter sound subtask, only one pronunciation for each letter is accepted, even though some letters have two or more correct pronunciations. We are keeping this observation here because the practice continues in Cluster 3 and thus remains an issue. Again, this issue was discussed in the October 2014 observation. We are repeating this observation because we believe this practice poses a threat to the validity of the test that should be mentioned in the report.

iv. English vs. local language pronunciation: The u and c sounds are different in English and local languages. The instructions for the letter sound task explain this; however, the child might not remember or understand when the assessment begins. Recommend: When providing the child with stimulus, repeat these instructions when giving the child the stimulus, ensuring that they understand that there is a difference. RTI response: Will check on this.

c. Timed reading task word difficulty. The first word of the reading task is Apeikolonjit. This word may be too long for some readers, and may stop them before they can proceed to other words that they might be able to read. Recommend: Change to shorter, simpler word to begin task.
d. Listening comprehension task - Question 3 asks (inferential question): Why was her aunt thankful? One answer is given: She chased the monkeys away. In order to avoid confusion and arbitrary decision-making on the part of the assessors, it is advisable to provide assessors with the range of possible acceptable options. Recommend: Give other possible responses should be listed – e.g., she threw stones at the monkey to chase them away, she saved the maize, she got rid of the monkeys. RTI response: This will be done.
e. Writing task. This year's EGRA includes a new task testing children's writing ability, in which they are asked to write several letters. Assessors were instructed to accept mirror image letters as correct - e.g., if the assessor says "b" and the child writes "d," it is to be marked as correct because it contains the stick and circle of the letter). There are two problems with this: (1) though the construct presumably being measured is a child's ability to write the letter, and not to know all the letters of the alphabet per se, some knowledge of the alphabet is needed in order to respond to the task. For example, if the child is asked to write the letter "b" and she writes "s," it will be wrong. Why then is "d" acceptable for "b"? (2) Though mirror images are acceptable (left to right), flipped letters are not (upside down). In this case, if the child is asked to write "m" and writes "w," given the logic of mirroring, why would this be marked as incorrect? Recommend: Provide rationale for using this procedure. If a rationale is not available, (1) mark a response wrong if a child writes the wrong letter (e.g., prompt=s, child writes t). (2) If mirror images are accepted, also accept upside down letters; (3) To avoid confusion, eliminate b/d (mirror), w/m (upside down), etc. because we can't be sure if the child knows the letter or not. RTI response: This will be discussed with HQ.

## Student interview

Some Problematic words/phrasing were noted.

a. Shin is similar to *ashina,* the Ngakaramajong word for *ass.* The use of words with unrelated cultural references can bias results. Recommend: This item be changed. RTI response: This will be checked.
b. Tongue means *language* and *tongue* in local language. When asked "Which language do you use at home?" a learner pointed to his tongue. Recommend: Reword. RTI response: This will be checked.
c. Q: Is there any day you did not come to school last week? Is this a double negative in local language? If so, it could lead to confusion. Recommend: Change to something like "Did you miss any school last week?" RTI response: This will be checked.
d. Q8: What language did your teacher teach you in at preschool or nursery school? Tangerine accepts only one option, yet teachers often teach in multiple languages. Recommend: Change to *What language did your teacher use* ***most*** *in preschool/nursery school?* RTI response: This will be checked.

## Field practice

a. Assessor marking not always noted. During school practice, some observers were noting the procedure but not always what the DQA was marking as correct or incorrect. Recommend: Ensure that DQAs observe how the assessors are marking the answers.
b. Limited practice with interview instruments in schools. The field practice focused on EGRA. In some cases, assessors used some of the interview instruments; however, most were presented during day 9 of the training, with no opportunity for assessors to use them in the field. This omission means that assessors will not get real-context practice using the instruments, and that the opportunity to identify problems with the instruments based on real-life practice (e.g., translation, clarity) will be lost. Recommend: In future trainings, ensure that assessors have an opportunity to use both EGRA and all interview tools during school practice.

## Debrief after field practice

a. Order of feedback. For the debrief session after school practice, DQAs and project staff gave their observations first, then asked assessors if they had any questions or additions. Adults learn best when the learning process starts with their experience and what they know. Moreover, related to the recommendation concerning experiential learning above, allowing learners to begin with an experience provides greater opportunities for learning, as opposed to starting learning with someone else telling them what they need to know. Recommend: Allow assessors to share their perspectives and questions before DQAs and project staff give their feedback.
b. Opportunity to enrich reflection on practice. Recommend: For the debrief session after practice in schools, ask each assessor to jot down personal impressions, concerns, or questions before the debrief, then allow each assessor to share in debrief. Also, consider collecting assessors' reflections and using this as content when all assessors come back together in plenary to lead general discussion.
c. Report on debrief: Recommend: Write debrief suggestions/decisions (e.g., on flipchart paper) during debrief to be kept for future discussions and documentation.

## Supervisors training session

The supervisor training was basically a guided reading exercise to convey the information in the Field Manual. No information was presented via video projector. While the Manual provided some useful information, its focus was on the C1 and C2 EGRA (e.g., administration of EGRA on paper and Tangerine, troubleshooting clues, Observation checklist) rather than taking into account information needs specific to C3 EGRA such as the new tasks (Orientation to print and letter writing). Their administration was not discussed, nor were they included on the assessors observation checklist. RTI shared all this information during the assessor training, but did not incorporate it into the manual (which nevertheless indicated it had been edited February 2015). Recommend: Any time tasks are revised or new tasks are added, that these changes be reflected in the Field Manual and observation checklists.

Table 2. Session procedures and comments

| Session | Procedure | Comments |
|---|---|---|
| TUESDAY 10 FEB | | |
| 8:30-10 Intro to program, importance of EGR: Lydia N, Tracy | PPT | Fine |
| Overview of EGRA and subtasks: Stella K, Lydia N | PPT | Fine<br>A couple edits:<br>Slide: What does EGRA measure?<br>*Fluency: delete "and proper expression"*<br>*Slide: What does EGRA measure: comprehension: delete "actively engage in"* |
| Results of EGRA: Tracy B | PPT | Fine |
| Tea | | |
| Learner agreement and rapport: Peter M | Form projected (in Word); facilitator explains<br>Video: Simulation with peers<br>Pair discussions | See Other recommendations #1 |
| Sound letter knowledge: Deborah | Form projected (in Word); facilitator explains<br>Video: Simulation with peers<br>Video: woman modelling pronunciation (showed twice)<br>Simulation: facilitator shows stimulus on PPT and has participants say the letters | *Q in video pronounced "qw": correct* |
| Lunch | | |
| QA: Participants ask questions | | |
| Vocabulary in English | Simulation: participant & trainer | |
| Letter sound identification: Lydia | PPT<br>Energizer<br>Group discussions: English vs. local language: what are the differences? E.g., Ng, Ny | |
| Syllable segmenting: Rosette | Form projected (in Word); facilitator explains<br>PPT | |
| (left at 4 pm, session lasted until 5) | | |
| WEDNESDAY 11 FEB | | |
| 8:30 Listening comprehension: Stella | Qs on the day before<br>PPT: in your language groups, take this time to examine your listening and questions.<br>Groups discussed<br>MN: a bit unguided: what should they be looking for?<br>Video: Simulation<br>Processing: What did you notice? MN: This is better than the way videos were | *Issue: Q3: Why was her aunt thankful? She chased the monkeys away. MN Other possible responses should b listed – she threw stones at the monkey to chase them away, she saved the maize, she got rid of the monkeys…)* |

| Session | Procedure | Comments |
|---|---|---|
| | used yesterday. Still, responses could have been guided toward learning points, and written somewhere for validation and future reference.<br>Simulation in MT<br>Processing of simulation: MN<br>Well done; good questions, good responses | |
| 9:32 Subtask 6: Letter writing: Lydia | PPT: Guidance & examples – e.g., nearly completed circle (pix)<br>Simulation<br>Processing<br>Review of protocol (from assessor's guide) | *Is there any way to confirm assessors' ratings? E.g., have them keep papers with student codes, review a sample*<br>*Slides: Suggest adding counterexamples – e.g., various examples of m given with two peaks – pointed or rounded – what about an attempt they would mark wrong? Suggest putting examples of acceptable and unacceptable letters in assessors' guide*<br>*Problem: If learner writes a mirror image, it is correct – NO! E.g., b is not d; B is not backwards B (interesting: if learner writes upside down or sideways, it's incorrect, as in m > w – if this is the case, why say no to w but yes to d?*<br>*MQ: In simulation, child wrote all letters. Should assessor rate each letter when written, or all at the end?* |
| 10:50 Orientation to print | PPT<br>Group work/discussions<br>Simulations: 1 on 1 in groups | |
| 11:35 Tablets: Rehama, Program Staff | PPT: Intro to tablets<br>Distribute tablets and have assessors put their names on the covers<br>Introduce the tablet | |
| Lunch | | |
| 2:20 pm Tablets continued | Project Tangerine and walk through each page, asking questions along the way<br>Pair practice with observations | *Same comment: asking questions along the way, but for the most part, pedagogy=telling rather than eliciting*<br>*Questionnaire, Q2: Do you speak Lhukonzo at home most of the time? Add follow-up Q: If not, what language do you speak? (with response options) – to get additional* |

| Session | Procedure | Comments |
|---|---|---|
| | | *info and to verify the child understood the question).*<br>*Same for next question.*<br>*Observation: During simulations, observers would say "you said /buh/ but it should be /b/." Suggest video recording simulation, showing both student with stimulus and assessor marking tablet. Then use video stop-action to have participants identify where assessor made errors.*<br>*Problem: What if student jumps around during letter sound recognition? Assessor needs to be able to follow, mark previous ones incorrect, etc. Tracy said that's true but this doesn't happen often.* |
| THURSDAY 12 FEB | | |
| 8:30 Tracy: summary from previous day | PPT: Leads discussion on various topics – e.g., when to press the start button – i.e., don't press until child begins<br>PPT: Research ethics, validity and reliability<br>PPT: Inter-rater reliability<br>Photos of judges with different scores (sports stadium) and same scores (Dancing with the Stars); criteria for judging dog shows.<br>Exercise: Everyone hops when I say the name of an animal we eat, don't jump if we don't. | *Effective: Projecting assessor's view (e.g., letters) and simulating the task, with a participant playing the child, and showing on the screen how to mark mistakes. (Evelyn tells me this was a recommendation from NORC.)* |
| 9:30 Rosette Overview of different tasks | PPT: Summary of tasks; orientation to print | *Why are we reviewing this today?* |
| 10:05 timing responses: 3 seconds, 5 seconds | Having plenary clap at 3 and 5 seconds | *Effective* |
| 10:30 Quiz | PPT: See questions at end | |
| 11:00 Review English letter sounds | PPT: trainer leads plenary in pronouncing English letter sounds | *Problem: E=ay (is this always the case in Ugandan English?)*<br>*Clipped sounds: Problem: They are pronouncing /q/ /qu/*<br>*Problem: Trainer should put microphone near throat; in front of her mouth, /t/ sounds like spitting* |
| 11:30 Guided practice, English letter sounds | | |

| Session | Procedure | Comments |
|---|---|---|
| 12:20 IRR: English letter sounds | 2 trainers demonstrate, participants mark on their tablets | |
| Lunch | | |
| 1:15 Guided practice, all tasks | Facilitator gives ive instructions<br>Corrected quizzes returned | |
| 3:30 IRR English | | |
| 3:55 IRR: Local languages | Simulation, participants note, upload<br>Observed Ngakarimojong group | |
| FRIDAY 13 FEB | | |
| 8:30 Plenary review of timed reading task. | Project Tangerine task and simulate reading, showing how to mark words wrong, mark the last word read. | *Effective* |
| 9:00 Agenda posted | | |
| 9:05 Results of IRR exercise, English | Projected 2 Excel tables in plenary:<br>1. Results by item: Column A=participants' names, B-X=labeled by item (e.g., letter A, c, T…), column filled in green if 90% or above, yellow if 80% or above, red below 80%<br>2. Results by assessor: Column A=participants' names, B=cells filled green, yellow, red for score, with scores in cells<br>Projected stimulus with problematic letters colored | *Very effective* |
| Results of IRR by language group & guided practice | Results of IRR presented to each language group while other groups continue to work on subtasks with DQAs leading sessions. | |
| Learner's context interview | | |
| Practice rapport building | | |
| Practice with IRR | | |
| Discussion of field practice | | |
| SUNDAY 15 FEB | Travel to Moroti District | |
| MODAY 16 FEB | Observations of field practice | |
| TUESDAY 17 FEB | Return to Kampala | |
| WEDNESDAY 18 FEB | | |
| 8:30 Debrief: School practice | Was the experience different from what you expected?<br>Tracy passes mike, participants ask questions, | |

| Session | Procedure | Comments |
|---|---|---|
| | Tracy answers – e.g., what if student starts to just give answers?<br>Areas for practice (PPT) – e.g.,<br>- 3-second rule for moving learners along<br>- What to do if a child remains silent?<br>- Listening story read too fast<br>- Eye contact with learner<br>- Adding extra words and support | |
| | Projects Excel file of data uploaded from practice lessons –<br>- How long?<br>- What time did they start? | *TOO SMALL!* |
| 9:05 Review of letter sound exercise | Project Tangerine letter sound task, simulate how to say "please go on" | |
| 9:30 Guided practice | Work in language groups reviewing results uploaded from field practice | |
| 2:15 IRR, English | Review of results from school practice | |
| 3:00 IRR, Local language | Groups sit by language and DQA reads results from practice session – e.g., Opolot spent too long on letter sounds, fine on timed reading, etc. DQA then tells them what to do: you look at the child, say go on… | *This feels very didactic and not very effective. As a learner, how do I know what signals "too slow"? How do I improve? Better to video tape simulations in plenary, then through stop-action, debrief and take questions, writing rules on separate sheet to hand out later – one populated as much as possible by participants' own thoughts.* |
| THURSDAY 19 FEB | | |
| 8:30 Recap | | |
| 9:00 IRR results | English + local languages in groups<br>Participants worked in groups for guided practice | *Problem letters identified in IRR: H, n, f, l, e, r*<br>*Tracy announces that assessors should upload pronunciation tape onto tablets for review* |
| 9:30 IRR results: Local languages | | |
| 10:00 BREAK | | |
| 10:30 Practice | | |
| 11:42 Head Teacher interview | Project Tangerine tool; assessor reads instructions out loud<br>Facilitator walks through each page (projected). | |

| Session | Procedure | Comments |
|---|---|---|
| 12:30 Sampling | | |
| 1:15 Lunch | | |
| 2:15 Practice assessment on paper | | |
| 3:00 Field work code of conduct | | |
| 4:30 Deployment planning | | |
| **20 February 2015-Supervisors training** | | |
| 8.30. Welcome | Introductions, | |
| 8.40-10.00. Walk through field work manual<br>10.00- Tea break<br>10.30. School summary sheet<br>11.00. Last minute guidance and questions<br>11.30. Onwards to prepare for fieldwork, marking envelopes, learner sample slips, upload tablets and packing | Reviewing Fieldwork manual; Supplied copies to Supervisors. Read in turns.<br><br>Supervisors asked questions majority were responded to.<br><br>Supervisors were engaged in preparing field materials. Writing details on school envelopes and writing learners sampling cards.<br>. | *The session did not have projected presentations, it depended on guided reading. Presentations were not systematic.*<br><br>*There were contradictions on how to pick learners from class. Some DQAs said, each Assessor should pick his/her learners from class, others suggested, let learners be picked in sets of eight. There was no concrete decision on this.*<br><br>*According to the manual and training discussions, supervisors were assigned many roles that seem unrealistic to perform. They are core Assessors for the learners because they are better skilled, have to oversee work of Assessors, undertake teacher and head teacher interviews, pick learners from class. It is not very clear how these people will execute all that.*<br><br>*Distribution of Field Manual during Supervisors training does not serve a good purpose. The Manual seem to be useful to Assessors too. It should be part of training package provided to all at beginning of the training.* |

# ANNEX 2: OBSERVATIONS FROM FIELD VISITS DURING CLUSTER 3 BASELINE DATA COLLECTION

**MEMO TO:** RTI/School Health and Reading Program (SHRP)
**FROM:** NORC/Panagora Performance and Impact Evaluation (P&IE) Team
**DATE:** March 12, 2015
**SUBJECT:** Performance feedback on observed SHRP activities in February 2015

## EXCERPT FROM MEMO RELEVANT TO OBSERVATION OF FIELD WORK

### Cluster 3 Field Data Collection, February 24-27, 2015

**Purpose:** The main activity was collecting EGRA data from P1 learners along with accompanying information from teachers and school administrators.

**Content/Approach:** Assessors conducted EGRAs with thirty P1 and ten P3 students, along with a learner context interview, teacher and head teacher interview, and inventory of school assets in each school. Classroom lesson observations were also conducted in a subsample of schools. The DQA observed classroom lessons and the rest of the data collection was performed by Assessors and Team Supervisors. Each Local Language group had three teams of three Assessors and one Supervisor, who reported to the DQA; each team conducted EGRA in one school. The DQA was the overall supervisor for the three teams, and worked with a different team every day.

**Relationship to work plan:** The outputs for this activity correspond with Intermediate Result 1.5, Programs and Policies informed by data and research.

### Appreciative feedback

***EGRA Field Data Collection***

- **Teams arrived at schools on time.** Teams arrived at the schools early in the morning, with enough time to prepare the assessment area, sample the learners, before beginning the assessments.
- **Teams followed field protocols while at the schools.** As specified in the field protocols, upon arrival at the schools, Assessors first went to the head teacher's office to introduce themselves and request permission to carry out EGRA activities. They also respected the teachers and learners they worked with.
- **Teams established good rapport with students.** Realizing that good rapport with students made them more likely to consent to participating in the assessments, Assessors worked to build a good connection with the students upon first meeting them by taking time to chat on various topics of interest. This initial rapport also helped to build students' confidence to respond to the assessments.
- **Assessors followed instructions outlined for each task.** The majority of the Assessors followed and clearly read to the students the instructions provided on the tablets.

- **Assessors were competent using the tablet technology.** All Assessors were able to use the tablets: they easily scrolled through different tasks, began and ended tasks, saved interviews, and created new interview forms for different students.
- **Assessors well conversant with the area local language.** Assessors spoke and read the local language fluently, and easily interacted with learners using the area local language.
- **Lusoga teams prepared an appropriate sitting arrangement for the assessments.** During assessments, Assessors and students sat opposite each other on different seats, which encouraged good eye contact and offered ample space and good positioning for reading and writing during the assessments. Assessors also had space to place their materials and keep the tablets from view so as to not distract students during assessments.
- **Assessors familiar with stimuli to use on each task.** All assessors provided the correct stimuli to students during the assessments.
- **Assessors practiced examples provided for the different tasks.** On all tasks with examples, Assessors demonstrated the examples as they appeared in the instructions and did not take shortcuts.
- **Assessors marked the last item attempted on timed tasks.** During letter sounds and reading tasks, Assessors were mindful to mark the last item a learner attempted.

**Constructive Feedback**

***EGRA Field Data Collection***

- **Interview environment was not adequate to ensure privacy and learner concentration.** Based on previous EGRAs, school compounds are the most available spaces for conducting student assessments. Unfortunately, these are often near main roads and pathways, or close by other Assessors conducting assessments. In some schools, the waiting area for students was in the same room as the assessment, and students were able to observe their peers being assessed. Assessor training should cover issues of location when conducting EGRA in schools and help Assessors identify and/or organize assessment locations with the least distractions and the most privacy. These issues were particularly problematic in Kakombo Primary School in Iganga District and Nandere Primary School in Kibuku district.
- **Inadequate sitting arrangement observed in Lugwere Local Language teams.** The two Lugwere Local Language teams opted for a seating arrangement where students sat on the same side of the desk as the Assessor, which discouraged eye contact and made it difficult for the Assessor to hold the student's attention throughout the assessment. In addition, Assessors placed assessment materials in positions that were inconvenient for the reading and writing exercises. For example, items the students used during the English vocabulary and letter writing tasks were placed next to the student on the same seat, where it was difficult for the student to read, point, or write as instructed, and most students felt uncomfortable moving the materials to a more comfortable (such as their laps or the desk). Currently seating instructions are left to the discretion of DQAs and this topic is not covered in Assessor training. There should be overarching guidance for DQAs on optimal seating and the topic should be included in Assessor training.
- **Replacement students picked after the morning sampling process were not randomly selected.** Teams followed the sampling methodology adequately but they were limited to selecting from only two replacements. In one of the schools, the two replacements picked were from same

stream. In addition, if teams encountered students who declined to participate and replacements were already used; Assessors picked any student from the class. To select these extra replacements, some teams used a convenient sample with the help of the teachers instead of using the prescribed sampling strategy. To avoid inconsistency in replacement sampling strategy, in the future, the number of replacements should be increased to ensure a random selection process.

- **Inconsistencies noted in administration and marking some tasks.**
    - *Inconsistencies in tapping the start button for all timed tasks.* For timed tasks, such as letter sounds and reading comprehension, some assessor began the timer before providing the prompts to students, and some began the timer after providing the prompts. This difference in assessment administration creates unfairness and can affect results.
    - *Differences in conducting syllable segmenting task.* We had noted in past observations that instructions regarding the number of times words are read to students seem to be confusing to assessors. When Assessors provide examples to participants, the instructions on the tablet shows words to be read once to the students; but later in the instructions, Assessors are directed to read the word twice. This creates confusion in administration of the assessment, and some Assessors read the word once throughout, some read the word twice throughout, and some varied the number of times they read the word. This difference in assessment administration creates unfairness and can affect results.
    - *Differences in handling* English vocabulary subtask on body parts.
        - Assessors practiced with two examples of English vocabulary: "nose" and "head," and asked students to point to the correct body part. Some Assessors did not correct students when they did not point correctly, but some did. Currently there is no instruction for Assessors to either correct or not correct students when they fail to identify the correct vocabulary word, and this is needed for consistency, fairness, and accurate results.
        - With the "nose" and "head" examples of English vocabulary, some Assessors prompted students by either providing the word in the local language or repeating the word in English, but some did not provide any prompts on marked learners as incorrect throughout the assessment.
    - *Differences in handling English vocabulary subtask on, identification of items in the room*. Assessors placed items such as books, pencils, paper, or erasers at the end of the desk furthest away from the student where it was difficult for the student to see them then asked them to point to or touch the items. When asked why the items were placed so far away, an Assessor said that was the instruction given by the site supervisor, and added that it is cheating on the assessment to place the items nearer to the learner. This situation needs to be remedied.
    - *Confusion when marking English vocabulary sub task about spatial words.* This subtask is highly subjective, requires critical judgment to decide on marking a learner correct or incorrect. As was noted in past observations, the most problematic spatial words were putting a pencil "on the paper" and "in front of you". "On the paper" was number one item assessed, and it was difficult to mark it adequately. Some learners were just dropping the pencil on the paper that was directly placed in front of them. Some continued with same action throughout, but some Assessors marked this particular item correct and others marked it wrong. It was the same case with marking item directing a learner to put a pencil in front.

Some learners held the pencil at the front without action, some were marked correct and others incorrect. Future training should provide more information on improving judgment and marking the above mentioned spatial words.

# ANNEX 3: EXAMPLE OF APPRECIATIVE AND CONSTRUCTIVE FEEDBACK IN A MONTHLY MEMO (RTI RESPONSES IN ITALICS)

We provide the following illustrative example of appreciative and constructive performance feedback provided during the reporting period, including RTI's response (in italics). Our resident team observed four separate teachers Early Grade Reading trainings on C1 P1/P2/P3, C2 P1/P2, and C3 P1 materials at Shimon, Kabulasoke, Bwera, and Canon Apollo Kabarole PTCs:

*Appreciative Feedback:*

- **Training stations were adequately staffed with administration and training personnel.** Training stations had a team of supervisors comprised of a Site Manager, MoES supervisors, externally sourced persons, and College Principals; and were supported by two program Field Assistants. These staff members supervised training sessions and provided feedback to trainers. In addition, every training station was supplied with an adequate number of trainers. With exclusion of Shimon CPTC, stations had 3-4 trainers allocated to a training group.

- **Training used appropriate teaching methods.** Most training methods used were appropriate and engaging, with session reviews, demonstrations, reflection, group activities, and plenary discussions.

- **P2 and P3 training sessions well facilitated.** Generally, P2 and P3 were well facilitated at training stations. Trainers followed the training guides' step by step procedures, explained classroom lesson teaching procedures and steps, and helped participants to understand the link between teachers' guides and corresponding pupil books. This enabled more robust plenary discussions with more learning and knowledge exchange. Most trainers had facilitated three or more sessions previously, and therefore brought a lot of experience and familiarity to the exercise.

- **Orthography sessions were relevant and useful to participants:** The morning Orthography sessions were of great benefit to the training participants. They asked questions to understand the local language alphabets, writing styles, and other unique rules. Many of the questions asked were seeking support in understanding basics of their respective local languages. Some teachers said this was the first time they had attended such a comprehensive local language lesson. The basics in the orthography lessons were useful during other local language training activities.

- **Time management:** Time was well managed at all training sites. Trainers respected the time allocated to training sessions and breaks.

- **Observed improvements in delivery of English Literacy sessions:** We observed English Literacy at Canon Apollo Core PTC Kabarole. Trainers using P2 materials were more competent delivering English Literacy sessions as compared to similar sessions observed in the previous years. They shared sessions amongst themselves, easily followed the training guides, presented practical and well thought through micro teaching lessons, and responded to participants' questions with confidence.

- **Training offered opportunities to participants to understand application of skills taught.** Training sessions were accompanied with demonstrations, plenary discussions and group work assignments for participants.

- **Active participation of leaders and consultants in education.** MoES officers, NCDC officers, College Principals, and External Education consultants were involved in the training. These officers attended training sessions, providing useful input during sessions, and provided welcome feedback to trainers, who felt privileged to receive feedback from education experts.

*Constructive Feedback:*

- **P1 material training sessions were overly dependent on video. Most trainers relied on** a video to demonstrate skills. Unfortunately, at all stations we observed, the video was in a Northern language most participants didn't know and therefore had had difficulty following even with the English translations. The video was missing some steps in the teacher guides, e.g., steps covered in terms 2 and 3. At Bwera PTC, training was halted during an equipment failure when the video could not be shown. Because equipment was shared, participants and trainers often had to move back and forth from training rooms to the room with the video, at times delaying training or forcing trainers to rush through materials to keep pace with the video schedule. While video could be very useful in the training sessions, there are several issues that need attention for it to function seamlessly.

  *RTI Response: In our estimation, the videos showing the EGR methodology were a huge success. They were not perfect but we felt they were so powerful that we "fast tracked" getting them to the teachers for the January training. In so many cases, this was the first time that Ugandan teachers saw a very good example of any type of lesson in an actual Ugandan classroom. The videos were sub-titled so others could easily follow. It is not possible nor is it the plan to make the videos for all languages. A higher priority is making videos for additional grades.*

- **Orthography sessions need fine-tuning.** All language groups were combined for the Orthography sessions in one room, which created a larger group than could be managed by 1-2 trainers. Participants were eager to ask questions, but many were not able to given the size of the group and time available. Some questions were not adequately responded to because the trainers were not experts in the training local languages. These issues could be addressed by reducing the size of the session and having experts available in all local languages represented.

  *RTI Response: Please clarify that all language groups were combined. The language groups should be homogeneous. As the training has evolved, more and more time has been spent developing and implementing orthography sessions – including master training for the language experts.*

- **There was little team/collaborative training observed.** Trainers who were supposed to be team teaching often instead used training sessions to prepare other lessons, read participants' reflective journals, and read newspapers. This was very distracting and discouraged collaborative teaching. The only collaborations observed were at Canon Apollo PTC where P2 material trainers were paired to facilitate a session, however, they were also not supported by other pairs. Future trainings should devise ways of promoting teamwork among trainers.

  *RTI Response: Training is hard work. Sometimes it is fine for one teacher to lead while the other is doing other supportive work. Of course, reading newspapers or not being engaged is not acceptable. But we do believe that the trainers work hard over the course of the weeks or even weeks. Of course,*

*the program has recognized now that we have a bigger cadre of trainers that we can be more selective when choosing our trainers.*

- **One training site had more trainers than needed, but trainers did not cover all local language requirements.** Shimon CPTC training site had 5-6 trainers per 40-45 participants. In some cases all the trainers were actively involved in leading and supplementing sessions, but we also observed trainers who appeared to be extraneous and uninvolved. There were several trainers who were not conversant with the local language. The number and composition of trainers needs review, particularly with P2 materials, to ensure that the recruitment and deployment criteria yield trainers who know the local language.

  *RTI Response: There are many participants in a training, some of which may be the official trainers and others (such as University staff and MoES) there to provide support and, in the case of the MoES, visibility to participants of MoES "buy in" to the program. Of course, the main trainers need to know the local language and the program has strived to ensure this is the case.*

- **Some sections in teacher's guides were not discussed.** It appeared trainers were hesitant to discuss sections on differentiated learning and teacher's tips outlined in the teachers guides. Yet these sections contain important information that would address some of the frequent questions participants posed regarding application of the skills to big classes they handle in their respective schools.

  *RTI Response: It is true, sometimes there is information in the teacher's guide that may not be fully covered in the training. It is certain that no one would argue that there is too little content in the training, in fact, we often hear and believe the opposite – there is a tremendous amount of information conveyed to teachers in the course of the week. At the same time, the teachers guides are valuable resource tools for the teachers, and some of it can be referenced for them to read later. We are continually working to streamline the training content and helping the teachers to glean the most crucial information.*

- **Many P1 trainers facilitated sessions while reading directly from the books.** Although the teacher's guides have scripted lesson procedures, trainers should be familiar enough with the content beforehand so they are not directly reading the text, and able to provide some additional explanation. Directly reading without supporting explanations made these sessions much less effective.

  *RTI Response: Yes, this is still the case in some instances. In our estimation, teachers are becoming ever more confident and familiar with the material.*

# ANNEX 4: CHALLENGES TO THE IMPACT EVALUATION, AS PRESENTED IN SEMI-ANNUAL REPORTS PRESENTED IN JUNE 2013, OCTOBER 2013, APRIL 2014, OCTOBER 2014

1. Result 1: Delays in the implementation of Result 1 continued through October 2013. Although all the trainings have taken place, including refresher TOT and teacher training on Cluster 1 P1 materials (teacher guides and primers), these instructional materials were still being distributed to Cluster 1 schools as late as September/October 2013. Our understanding is that materials have not reached all schools at the time of writing this report. In addition, the original plans that included three different treatment arms were modified and treatment was uniform across all schools. Baseline data collection for Cluster 1 was completed successfully in February, and follow-on data collection for Cluster 1 is being fielded among a sub-sample of primary schools. These delays and modifications to the implementation do not pose serious risks to the evaluation at this juncture. We plan to evaluate the impact of the program as it was implemented.

   While the implementation changes/delays are not a risk to the evaluation design, an important fact to keep in mind, however, is that we do not expect to see the impacts of the full Result 1 intervention (teacher training and instructional materials) during this first impact analysis, using Oct/Nov 2013 data. However, the Oct/Nov 2013 data will provide us with an opportunity to measure the impact of multiple rounds of teacher training.

2. Result 1: The most recent version of the SHRP PMP indicates that no data will be collected from Cluster 2 in 2016. Going forward with this decision would imply that the impact evaluation for Cluster 2 would only be possible for P1 and P2 but not for P3. Given that Cluster 1 did not receive the full intervention in 2013, Cluster 2 will be the only group that will have a chance to receive three years of full treatment from the beginning of their primary education. The Evaluation Expert already mentioned this omission as a concern to USAID and to the IP as well.

3. Result 1: Data for the second EGRA wave are being collected as we write this report. Initial information from the field indicates low response rates (i.e. low numbers of students are being found) in the schools in the Central Region compared to baseline. We are currently working with the IP to try to address this problem and minimize the risks of having a small sample.

4. Result 2: There are several issues related to sample that have surfaced during the ongoing KAP data collection, which are likely to pose threats to the evaluation of Result 2 activities.

   We noted in our first Semi-Annual Report that, it was not possible to include boarding or partial boarding schools -very common among post-primary establishments- in the evaluation sample, given delays in obtaining parental consent for the KAP Survey during the school year. We decided, however, to use the second round of the KAP survey (KAP2) to collect additional baseline data from Cluster 1 boarding and partial boarding post-primary schools by distributing parental consent forms to students before the school break. The idea was to ensure that the

baseline survey consisted of a representative sample of post-primary schools, thereby allowing us to generalize the results of the impact evaluation to all such schools in the districts.

We recently learned of several problems that the IP is encountering with the supplemental boarding school component of the second round of KAP surveys. These problems could potentially have serious implications for sample size and the representativeness of the post-primary school sample:

- The IP faced resistance to data collection activities from some schools, where principals cited concerns that the survey would take away from exam preparation time (national exams in post-primary schools begin in the 2nd week of October) and some head teachers did not distribute consent forms to students at all. These schools could not be interviewed.
- Some schools closed before the end of the term and consent forms were not distributed on time. These schools could not be interviewed.
- Other programs related to HIV/AIDS have interacted with some of the schools and, therefore, head teachers decided not to participate in KAP. This is particularly the case of private secondary schools. These schools could not be interviewed.
- The sample frame that the IP provided NORC for selection of the school sample for the KAP2 contained errors; it included schools that already participated in the first round of KAP. In cases where these schools were randomly selected for the KAP2 sample, they had to be removed from the sample and, where possible, replaced.

NORC has requested from the IP a list of all schools in the KAP2 sample with disposition comments for each of the schools. After evaluating the situation we will have a clearer impression of the effect that these problems can have on the evaluation. At a minimum, we expect a reduction in sample size.

5. Result 2: As mentioned above, SHRP decided not to include post primary establishments in new treatment districts (Cluster 2 and after). Therefore, we will only be able to assess the impact of the Result 2 intervention on post-primary educational facilities for Cluster 1 schools.

6. Result 2: Based on the most recent PMP, we note that the Result 2 intervention will no longer be conducted in Cluster 3 districts and schools. As a result, NORC will focus its evaluation of Result 2 on Cluster 1 and Cluster 2 schools.

7. Result 1: Given program implementation delays in Year 1, the academic term was delayed for one week in the 11 districts of Cluster 1 where the IP is working in order to build in time to prepare and have teacher guides ready for the second training of teachers. Additional classes to compensate for the one week delay are not currently planned. An equivalent delay did not occur in the control district schools; therefore, the academic year in those schools will be one week longer. We do not anticipate a visible effect, but it is worth mentioning how the reality of the program may affect the evaluation.

8. Result 2: After NORC selected the samples for the impact evaluation of the School Health activity, the focus of the intervention underwent some changes in order to align with PEPFAR priorities. We were informed that the intervention would target large schools (with over 150 students) in high HIV prevalence districts; this brought into question the external validity of the impact evaluation and the ability to include non-intervention districts with similar characteristics to treatment districts in the design. However, these new criteria do not seem to have affected the actual selection of districts and we will proceed with the original evaluation design. However the number of treatment schools

increased. The IP went ahead with the selection of schools for treatment and control before NORC could approve the selection. As a consequence no replacements for control schools were selected. This can result in a smaller sample than needed. The Evaluation Expert discussed this issue with the IP and USAID.

**April 2014:**

3. Result 1: During the Cluster 2 EGRA training and pilot test, the P&IE team observers noted some issues related to the implementation of three specific EGRA subtasks – Letter Sound Knowledge and Word Segmenting, and Oral Passage Reading - and that could have negative implications for the impact evaluation. Annex 3 describes the issues in great detail and also lays out the implications for the impact evaluation. In short, SHRP was using very stringent requirements for accepting letter sounds as correct; for example, while the EGRA toolkit states that "For consonants that can represent more than one sound (i.e., c, g), either answer is acceptable. For vowels, either the short or long sound is accepted (/i/ as in pin or as in pine)," in the SHRP implementation of EGRA only one sound per vowel was being accepted as correct. As well, local pronunciations of words – e.g. "muzzah" for mother – were being marked as incorrect. This raises the concern that learners who actually know correct letter sounds are assessed as not knowing them, since trainers were instructed during training to mark as wrong any very slight deviation from the "ideal" sound of a letter.

   This approach can bias the assessment in favor of treatment schools, where students are being taught one correct letter sound or a specific pronunciation of a word, relative to control schools, where a broader set of letter sounds and pronunciations are being taught. We can take as an example the letter B3: the sound of letter B is /b/ or /buh/4. Both sounds are correct and accepted as building skills towards early reading ability. However, the current application of EGRA in Uganda only accepts a perfect clipped sound /b/ as correct. Marking /buh/ as wrong is likely to punish learners in control schools more than it punishes learners in treatment schools, because teachers in treatment schools are trained to teach /b/ as the only correct sound while teachers in control schools are likely to use either /b/ or /buh/ given that both sounds are considered correct. This approach of "teaching to the test" will bias impact findings in favor of treatment schools. NORC is exploring options for measuring this bias in order to adjust impact measures; towards this end, we briefly discussed some alternatives with USAID, such as measuring the bias by conducting experiments to test more and less restrictive versions of EGRA administration.

4. Result 1: Possible contamination of controls. Because the SHRP team is not planning to expand SHRP implementation to additional districts for Cluster 1, they are planning to implement Result 1 activities in control CCTs in the 11 original districts starting in 2014 in order to meet target numbers of trained teachers. However, they plan to exclude the control schools within the control CCTs which were selected for the EGRA data collection and intervene only in the schools from control CCTs which have not been included in the EGRA data collection. Hence, according to the SHRP M&E Team Lead, no teachers in any grade (P1 through P4) in the EGRA control schools will be trained; nor will instructional materials be distributed to these schools. CCTs associated with

[3] Similar problems exist with many other consonants such as D, T, P, K, G, etc.

[4] RTI International, EGRA Toolkit, March 2009
https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=149

these control clusters will be strictly instructed not to provide any assistance to these control schools.

Strict exclusion of control schools from treatment is critical for the integrity of the impact evaluation design. While SHRP staff has assured us that no control schools will receive any semblance of the Result 1 interventions, we are nonetheless concerned by the possibility of contamination through CCTs or spillover of materials. Any contamination of the control schools will lead to underestimation of the effects of the SHRP Result 1 interventions. We have made this concern clear to both the IP and USAID, and requested that SHRP put in place adequate safeguards to ensure that the control schools in our sample will not be contaminated.

5. Result 1: Non-systematic replacement of sample schools. During Cluster 2 baseline data collection in Mbale district, the SHRP team opted to exclude control schools that use or were presumed to use Luganda and English instead of Lumasaaba as the medium of instruction. The appropriate procedure to replace these schools (following the replacement rule provided) was not followed. Two of these non-Lumasaaba instruction schools were replaced by schools in which the medium of instruction is Lumasaaba; these replacements were picked from the list of preselected schools designated as replacements. The rest of the non-Lumasaaba instruction schools in the district sample were neither assessed nor replaced. We indicated to the IP and USAID that this approach was neither appropriate to keeping the integrity of a random sample nor conducive to comparing SHRP schools to the average public school in Uganda. First, replacing sample schools with hand-picked replacements creates problems with the sample balance. Second, the aim of the evaluation is to assess reading ability of learners in English and local language. While it is not possible to test them in the local language (Lumasaaba, in this case) in schools that do not teach in Lumasaaba, it would still have been possible to test student's performance in English. As such, NORC's Evaluation Expert urged SHRP staff to conduct the EGRA in English in these schools as soon as we learned of the situation. However, the SHRP team did not comply with this request in a timely manner. Therefore, NORC decided that the impact analysis will need to exclude Mbale district altogether.
6. Result 1: Manafwa district is encountering a serious crisis created by teacher transfers in the region. We learned during field observations that most of the teachers trained by SHRP in January 2014 in this region have been transferred to other schools: four of the treatment schools visited by our local staff did not have a trained P1 teacher, because s/he had been transferred. It will be critical to have information about the whereabouts of teachers trained by SHRP, since transfers of trained teachers away from treatment schools will have a severe effect on the impact evaluation. If these teachers end up at control schools, the impacts will be even more skewed. We will work with the IP and through our performance evaluation to try and capture the movement of trained teachers between schools.

**October 2014**

1. Result 1: Possible contamination in the control group. In October, we learned from RTI that Mango Tree Project was working in Otuke, a control district for the SHRP evaluation, and providing their literacy intervention to two control schools in the SHRP sample. This occurred despite careful coordination between Mango Tree and SHRP. The schools in question were replaced for others, however the replacement schools have no baseline and their usefulness is limited. In addition, it is possible that some contamination has already occurred as we do not have any type of control over Mango Tree activities in the district.

2. Result 1: Sample size adjustments between rounds. The IP has changed sample sizes of each cohort between rounds. For the first cohort of students (Cluster 1), the February 2013 baseline included 280 schools to allow for analysis of 3 treatment arms, controls in treatment districts, and controls in comparison districts. However, the Cluster 1 Round 2 data collection conducted in October 2013, RTI collected data only in a subsample of treatment schools (168 of the 280), since a decision was made to only focus on one (and not 3) treatment. For Cluster 1, Round 3 in October 2014, however, RTI reverted back to data collection from 280 schools to account for that fact that the 3 treatment arms were implemented in the second year. This use of unbalanced panels does not preclude us from conducting a rigorous evaluation; however, it makes the process less transparent and prevents us from having measurements year by year without loss of information and precision. A similar change happened for Cluster 2. In this case, RTI requested NORC to calculate a sample size large enough to be able to analyze results at the district level. At baseline, in February 2014, data was collected from enough number of schools to calculate impact at district level; however, in for the first follow up in October 2014, the IP decided that district level analysis was not of interest and reduced the data collection to a subsample of the original schools. Although we will not be able to say anything about impact at district level, if properly implemented, this change should not prevent us from analyzing results at language level. In general, NORC recommends following the original samples over time to produce a more streamlined and transparent process and more comparable results across year.